

Article

Deep Learning-Based Morphological Classification of Human Sperm Heads

Imran Iqbal ¹ , Ghulam Mustafa ² and Jinwen Ma ^{1,*} 

¹ Department of Information and Computational Sciences, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China; imraniqbalrajput@pku.edu.cn

² Department of Biomedical Engineering, College of Engineering, Peking University, Beijing 100871, China; mustafabme@gmail.com

* Correspondence: jwma@math.pku.edu.cn; Tel.: +86-10-6276-0609

Received: 15 February 2020; Accepted: 15 May 2020; Published: 20 May 2020



Abstract: Human infertility is considered as a serious disease of the reproductive system that affects more than 10% of couples across the globe and over 30% of the reported cases are related to men. The crucial step in the assessment of male infertility and subfertility is semen analysis that strongly depends on the sperm head morphology, i.e., the shape and size of the head of a spermatozoon. However, in medical diagnosis, the morphology of the sperm head is determined manually, and heavily depends on the expertise of the clinician. Moreover, this assessment as well as the morphological classification of human sperm heads are laborious and non-repeatable, and there is also a high degree of inter and intra-laboratory variability in the results. In order to overcome these problems, we propose a specialized convolutional neural network (CNN) architecture to accurately classify human sperm heads based on sperm images. It is carefully designed with several layers, and multiple filter sizes, but fewer filters and parameters to improve efficiency and effectiveness. It is demonstrated that our proposed architecture outperforms state-of-the-art methods, exhibiting 88% recall on the SCIAN dataset in the total agreement setting and 95% recall on the HuSHeM dataset for the classification of human sperm heads. Our proposed method shows the potential of deep learning to surpass embryologists in terms of reliability, throughput, and accuracy.

Keywords: classification; convolutional neural network (CNN); deep learning; infertility; sperm head morphology

1. Introduction

Human spermatozoon is the gamete—the male reproductive cell—that may fertilize the mature oocyte. It is produced in the seminiferous tubules of the testicles. Structurally, normal human spermatozoa have four main parts: head, midpiece, tail, and end piece, as shown in Figure 1 [1]. A normal human sperm has a smooth oval head which looks like the shape of an egg. The sperm head can be further divided into two subunits: nucleus and acrosome.

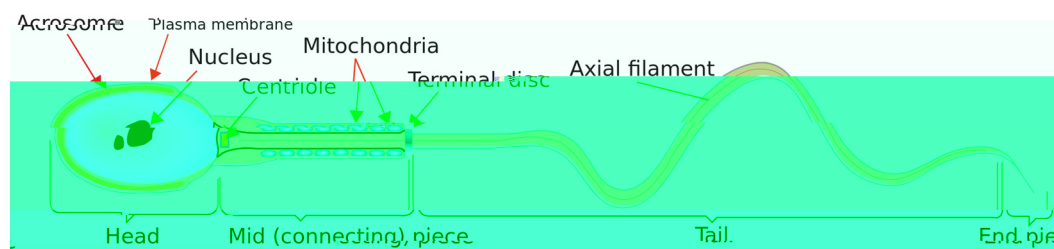


Figure 1. The diagram of a human spermatozoon.

Embryologists can observe the behavior of a spermatozoon by means of a microscope. It resembles a translucent tadpole since it has a long lashing tail and a circular head. The shape of the tail expedites the spermatozoon to progress keenly after it is evacuated from the reproductive gland. The tail supports propulsion of the spermatozoon towards the uterus in pursuit of an egg in the salpinges. Moreover, the tail of the spermatozoon enables the required motion to bind to and further penetrate a mature oocyte when it arrives.

Male human infertility or subfertility occurs when male reproductive cells fail to let a fertile female conceive a child or delay pregnancy after one or more years of regular unprotected sexual intercourse [2,3]. When a man fails to produce an adequate quantity of spermatozoa and/or produces low quality spermatozoa, these spermatozoa are called sub-optimal.

The generation of low quality spermatozoa minimizes the pregnancy rate [4]. These spermatozoa can be immotile or/and abnormal in shape. The immotile spermatozoa cannot move up to the fallopian tubes. As a result, they cannot fertilize a female ovum. The abnormally shaped spermatozoa may be able to travel, but even if they manage to reach the female gametocyte, they may not bind to and penetrate its shell and therefore the woman may reduce her chance of getting pregnant. Conversely, when a male body produces a low number of reproductive cells, the probability that one of the sperm in the semen unites an egg to form a zygote significantly decreases. There are some possible factors in the male body such as the age, anxiety, pathogens, and diet, which may impact the number of abnormal sperm in the semen [5,6]. It is clear that high sperm head deformities lead to low fertilization, implantation, and pregnancy rates [7].

Human infertility is a disease of the reproductive system that affects more than 10% of couples across the globe and over 30% of reported cases are related to men [8]. The crucial step for male fertility diagnosis relies on the examination of sperm morphology through the seminogram. The key types of defects of the abnormal sperm are: head, neck, tail and excess residual cytoplasm [9], but head abnormalities play a major role in male infertility. There are two main tasks in sperm morphology analysis; the first is to classify the types of defects in the sperm head, neck, and tail, and the second is to estimate the number of abnormal sperm. In this study, we emphasis on the classification of the head morphological defects or abnormalities.

In practice, the results derived from manual morphological analyses of sperm rely heavily on the expertise of laboratory technicians [10]. Moreover, this manual examination is laborious, non-repeatable, time intensive, and there is a high degree of inter and intra-laboratory variability [11]. For animal spermatozoon analysis, there exist certain computer-aided sperm analysis (CASA) using commercial software. However, human semen samples have a much lower quality of spermatozoa than animal semen samples [12], and thus the same software may not be directly applied to human spermatozoon analysis. Furthermore, it was found that the application of the CASA system to analyze human spermatozoa required human assistance which may affect results of the assessment subjectively [13].

According to the above analysis, it is important to design accurate, automatic, and efficient artificial intelligence (AI) systems to improve the numerical analysis of human spermatozoa from the sperm images. Actually, the morphological classification of human sperm heads plays an important role in the numerical analysis of human spermatozoa, which has already attracted extensive interest relating to the diagnosis of male infertility. Our main interest here is to focus on the development of deep learning model to extract features directly from sperm images for morphological classification of human sperm heads. According to the World Health Organization (WHO), there are 11 abnormal categories of human sperm heads, which are defined according to certain particular morphometric characteristics of the heads. They differ in shape, size, and texture in a very complicated way so that the task becomes extremely difficult even for a human expert. In addition to intra-class differences, there are also inter-class similarities. For instance, an elongated Amorphous head is similar to a

From the public SCIAN dataset [14] and recent studies, it was found that the morphological classification of human sperm heads is very challenging for the following reasons: (1) There is a high degree of inter-class similarities as well as certain intra-class differences in some cases; (2) Low-magnification microscopic images of sperm heads are very noisy; (3) The size of the images is very small: the length and width of the sperm heads are about 4 μm and 3 μm , respectively, and the size of each image is approximately 35 by 35 pixels; (4) The number of sperm head examples is insufficient for training a complex machine learning model; (5) The two-thirds of the examples in the SCIAN (partial agreement) dataset consists of only 2-out-of-3 human expert agreement; (6) The classes are highly imbalanced (e.g., the Amorphous class has ten times more examples than the Small class); (7) The Amorphous class has no common structure, and their forms can change in different ways.

The main aim of this research is to develop, implement, and calibrate an advanced deep learning model in the context of morphological sperm assessment. This specialized deep CNN architecture can accurately classify microscopic human sperm head images according to WHO criteria. Our proposed deep learning architecture is good to expedite the automatic classification process of human sperm heads. This innovative method has the potential of deep learning to exceed embryologists in terms of accuracy, reliability, and throughput.

2. Related Work

According to the guidelines of WHO, there are 11 categories of abnormalities of human sperm heads: Tapered, Pyriform, Amorphous, Small, Small acrosome, Large, Large acrosome, Round, Two heads, Vacuolated, and Vacuoles in the post-acrosomal region. Among them, the Tapered, Pyriform, Amorphous, and including Normal categories can mainly be discriminated by the precise shapes of their samples. Therefore, it is extremely challenging to distinguish them even by an embryologist. However, the remaining abnormal categories can mainly be discriminated by the different sizes of their heads or the existence of vacuoles or the acrosome and thus it is relatively easy to distinguish and recognize them. For sperm classification tasks, conventional machine learning algorithms have been adopted to alleviate the laborious work of embryologists and improve classification performance. Nonetheless, the input of these algorithms contain certain manually extracted spermatozoon features like the head perimeter, area, and eccentricity [15,16]. Although several approaches have been established for the semen analysis of animals (e.g., [17,18]), there are only a few approaches for the morphological classification of human sperm heads. We now briefly review some machine learning approaches related to the morphological classification of human sperm heads.

In 2017, Chang et al. [14] introduced a gold standard dataset, SCIAN-MorphoSpermGS, for the analysis and evaluation of morphological classification of human sperm heads. Notably, there had been no open and free available dataset before this gold standard dataset became public. The SCIAN dataset has five classes of human sperm heads for semen analysis namely: Normal, Tapered, Pyriform, Amorphous, and Small, which are available in the WHO laboratory manual. It consists of 1854 sperm head images, which were labeled by three Chilean referent domain experts as specified by the guidelines of WHO. Chang et al. [19] further proposed a two-phase analysis pipeline, CE-SVM, for the morphological classification of human sperm heads in the SCIAN dataset. In the first phase, a classifier is trained to distinguish the Amorphous category from the remaining four categories. In the second phase, four classifiers are trained for the four non-Amorphous categories, where each classifier aims to distinguish the specific non-Amorphous category from the Amorphous category.

From a different direction, Shaker et al. [20] released the Human Sperm Head Morphology (HuSHeM) dataset and proposed an adaptive dictionary learning (APDL)-based approach, which extracts certain square patches from the sperm head images to train the dictionaries to recognize those sperm head categories. At the evaluation stage, square patches are recreated with the dictionary and the minimum overall error among those of all the categories is computed to identify the best sperm head category. Recently, with the fast development of deep learning techniques, Riordon et al. [21] used a VGG16 architecture (FT-VGG) for the morphological classification of human sperm heads.

First, the VGG network was pre-trained on ImageNet [22] and then fine-tuned on the SCIAN dataset. Their experimental results demonstrated that this automatic deep learning method can facilitate and boost the seminogram effectively.

3. Methodology

3.1. Datasets Description, Partitioning, and Augmentation

SCIAN [14] is a gold-standard dataset for the morphological classification of human sperm heads with five categories: Normal, Tapered, Pyriform, Amorphous, and Small. The manual labeling of sperm head images in this dataset was independently performed by three referent Chilean experts who had experience in sperm morphology examination for several years. The images in this dataset are of greyscale with stained sperm heads, being taken at 63× magnification and their height and width are both 35 pixels or 7 μm. There are three separate agreement settings among three domain experts: no agreement, partial agreement, and total agreement. The first set consists of 1854 sperm head images (175 Normal, 420 Tapered, 188 Pyriform, 919 Amorphous, and 152 Small), but an image in this set can be labeled manually into three dissimilar classes by three domain experts. The second set comprises 1132 images (100 Normal, 228 Tapered, 76 Pyriform, 656 Amorphous, and 72 Small) but an image can be labeled into two different sperm head classes. The third set includes 384 images (35 Normal, 69 Tapered, 7 Pyriform, 262 Amorphous, and 11 Small), all three experts assigned the same class label to a sperm head image. From the number of images in these three sets, we can appreciate the difficulty of the morphological classification of human sperm heads even by human experts. For illustration (Figure 2), we show typical samples of human sperm heads of microscopic images of the five classes in the partial agreement setting of the SCIAN dataset and the four classes of the HuSHeM dataset.

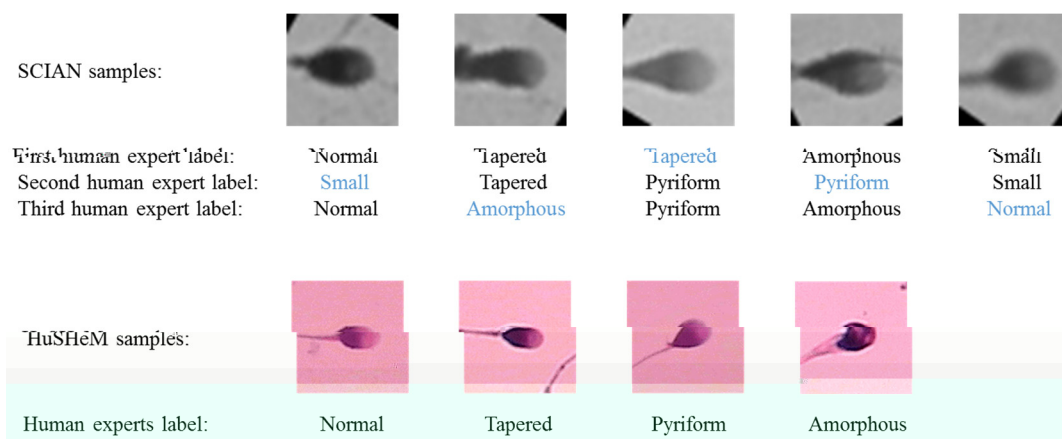


Figure 2. Typical samples of human sperm heads of microscopic images of the five classes in the partial agreement setting of the SCIAN dataset and the four classes of the HuSHeM dataset.

For effective usage of the SCIAN dataset, all images are converted into three channels and rotated so that all human sperm heads share the same orientation. For the convenience of comparison, we also adopt a stratified five-fold cross-validation scheme as used in [21]. That is, the SCIAN dataset is randomly partitioned into five parts, where the four parts that contain approximately 80% of the data from each class form the training set, while the remaining part which has roughly 20% of the data from each class forms the test set. The complete training/evaluation procedure is repeated five times for all possible choices of the training and test sets and the average results is reported. To compare the performance of our proposed model directly with the previous published results [20,21], each five-fold cross-validation procedure runs three times for stability. In addition, 20% of the fold-1 images are considered as the development set to tune the hyperparameters of our proposed network (see Table 1 for the details).

Table 1. The stratified five-fold partition of the SCIAN dataset (partial agreement), where the numbers denote the distinct sample sizes in different classes while the numbers in parentheses denote the total number of augmented with the addition of original samples in different classes at each fold. Moreover, the bold numbers in the training set (fold-1) denote the number of samples in different classes assigned to the development set for tuning the hyperparameters of the network. To avoid repetition, folds 2 and 3 are described together.

| Fold | Set | Sperm Head Classes | | | | | Total |
|---------|-------|--------------------|---------------|--------------|----------------|--------------|-----------------|
| | | Normal | Tapered | Pyriiform | Amorphous | Small | |
| 1 | Train | 80–20 (4860) | 182–46 (4896) | 60–15 (4860) | 525–131 (4728) | 58–14 (4840) | 905–226 (24184) |
| | Test | 20 | 46 | 16 | 131 | 14 | 227 |
| 2 and 3 | Train | 80 (6400) | 182 (6370) | 61 (6405) | 525 (6300) | 58 (6380) | 906 (31855) |
| | Test | 20 | 46 | 15 | 131 | 14 | 226 |
| 4 | Train | 80 (6240) | 183 (6222) | 61 (6283) | 525 (6300) | 57 (6270) | 906 (31315) |
| | Test | 20 | 45 | 15 | 131 | 15 | 226 |
| 5 | Train | 80 (6240) | 183 (6222) | 61 (6283) | 524 (6288) | 57 (6270) | 905 (31303) |
| | Test | 20 | 45 | 15 | 132 | 15 | 227 |

In order to tackle the issue of skewed classes and training image scarcity, we implement more augmentation options to the minority classes, and less augmentation options to the majority classes to balance the sample size in each class of the training set. Therefore, the training set is extended virtually for the deep learning task with the actual classes being balanced. For example, the Pyriiform and Amorphous classes in the fold-5 partition of the partial agreement setting (see Table 1) have 61 and 524 distinct images, respectively, but the sample image sizes in the two corresponding augmented classes are similar, i.e., 6283 and 6288, respectively.

As for the specific data augmentation, we adopt three common techniques for the SCIAN

Table 3. The stratified five-fold partition of the HuSHeM dataset, where the numbers denote the distinct sample sizes in different classes while the numbers in parentheses denote the total number of augmented with the addition of original samples in different classes at each fold. To avoid repetition, folds 1, 2 and 3 are described together.

| Fold | Set | Sperm Head Classes | | | | Total |
|------------|-------|--------------------|-----------|-----------|-----------|-------------|
| | | Normal | Tapered | Pyriform | Amorphous | |
| 1, 2 and 3 | Train | 43 (4730) | 42 (4620) | 46 (5060) | 42 (4620) | 173 (19030) |
| | Test | 11 | 11 | 11 | 10 | 43 |
| 4 | Train | 43 (4730) | 43 (4730) | 45 (4950) | 41 (4510) | 172 (18920) |
| | Test | 11 | 10 | 12 | 11 | 44 |
| 5 | Train | 44 (4840) | 43 (4730) | 45 (4950) | 41 (4510) | 173 (19030) |
| | Test | 10 | 10 | 12 | 11 | 43 |

3.2. Proposed Deep CNN Architecture and Learning Paradigm

according to the lowest loss of the model evaluated on the development set. Finally, the obtained model is used to assess the test set.

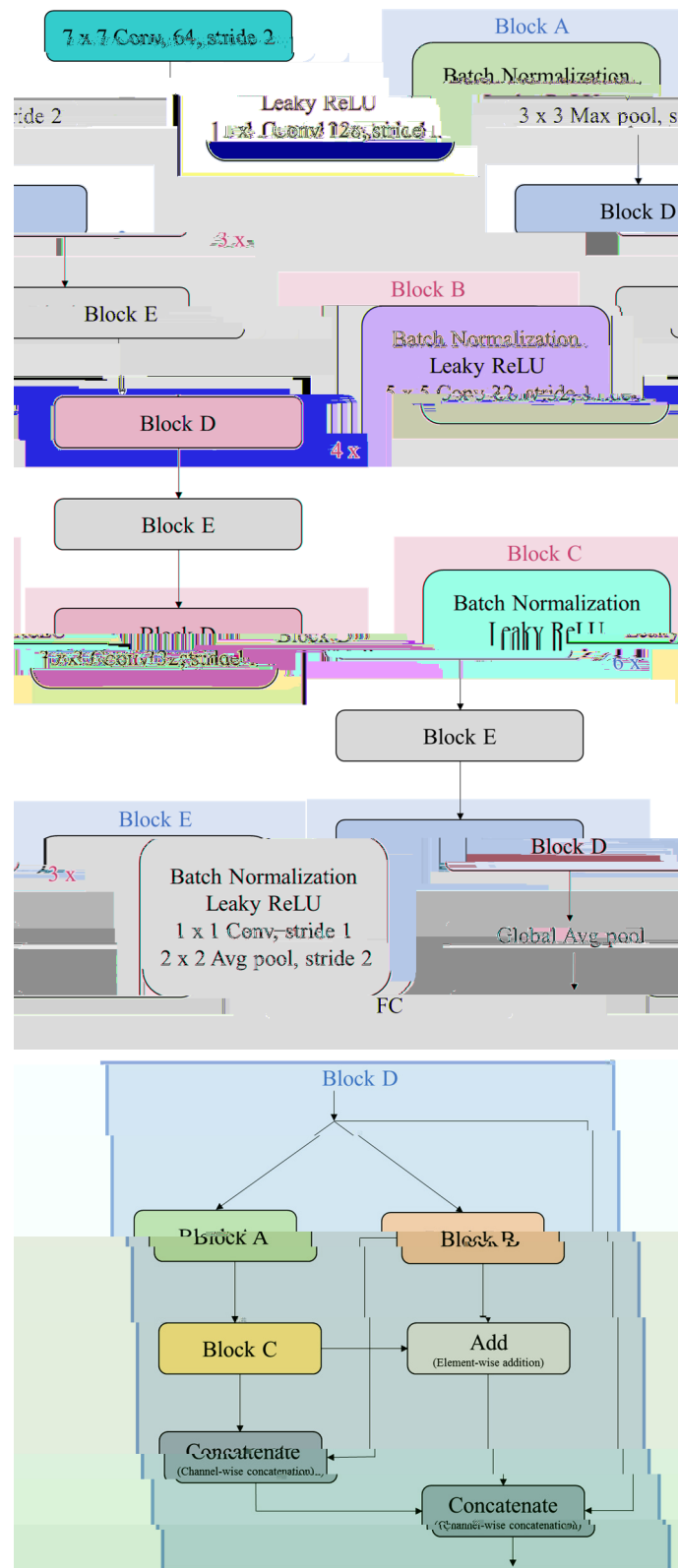


Figure 3. The layout of the proposed deep CNN architecture, where ‘FC’ denotes the fully connected layer.

4. Experimental Results

In this section, we ran five-fold cross-validation analyses for our proposed deep CNN model for the morphological classification of human sperm heads in the SCIAN and the HuSHeM datasets. We tested it on both partial and total agreement settings of the SCIAN and the HuSHeM, and compared our results with the state-of-the-art methods. We used the metrics of the precision, recall, specificity, F₁-score, Jaccard similarity coefficient, geometric mean (G-mean), Matthews correlation coefficient (MCC), and Cohen's kappa score (CKS) for the classification assessment and comparison. There are two types of averaging: macro-averaging and weighted-averaging. That is, when computing the average of the indices of the classes, equal weight is assigned to all the classes in the way of macro-averaging, while a different weight is assigned to a class that is proportional to the number of its images in the way of weighted-averaging. According to the stratified five-fold partitions of the SCIAN (in both partial and total agreement settings) and the HuSHeM datasets and the learning paradigm given in the previous section, we implement our model using the TensorFlow and Keras framework on a NVIDIA GeForce GTX 1080 card with 8GB GDDR5X memory. The training process takes roughly 18 h in total for the SCIAN dataset. We also evaluate our deep learning model on the HuSHeM dataset. The training process takes approximately 5 h in total on this dataset. In the following subsections, we summarize and discuss the experimental results and comparisons in both partial and total agreement settings of the SCIAN dataset as well as the results on the HuSHeM dataset.

4.1. On the Stratified Five-Fold Partition of the SCIAN Dataset with the Partial Agreement Setting

Our proposed model is first evaluated on the stratified five-fold partition of the SCIAN dataset in the partial agreement setting. We train the deep CNN architecture on each choice of training set in the partial agreement setting and tune the hyperparameters on the development set. The experimental results of our proposed model on the SCIAN dataset with the partial agreement setting is shown in Figure 4a–h. The detailed experimental results are shown in Supplementary Materials (Figures S1–S8). Figure 4a–b show typical classification accuracy and cost curves with the number of epochs on a specific choice of training and test sets. It is seen that the training process converged within 50 epochs. Notably, our proposed model achieves much better accuracy and recall than the previous methods in the partial agreement setting (Table 4). By the stratified five-fold cross-validation, we get the confusion matrix (Figure 4c), from which we can see how often images of each individual class (Normal, Tapered, Pyriform, Amorphous, and Small) are predicted by our proposed model on the test set in the partial agreement setting only for a typical run. We also get the average confusion matrix over 15 runs (5 folds × 3 runs) as shown in Table 5. After carefully examining these tables, we find that the Amorphous class is very difficult to distinguish from the remaining classes. The main reason for this may be that the Amorphous class has a variety of forms. On the contrary, we also find that the average true

recall, specificity, F_1 -score, Jaccard similarity coefficient, G-mean, ROC-AUC, PR-AUC, MCC, CKS, and evaluation time. The standard deviation in the last row of this table shows the stability of the result of our proposed model with a training run for each index. Since all standard deviations are less than 0.09, our proposed model is therefore quite stable with the learning algorithm.

Figure 4. The experimental results in the partial agreement setting of the SCIAN dataset: (a,b) Typical classification accuracy and cost curves with the number of epochs during the training on the training and test sets; (c) The confusion matrix on the test set; (d–f) The precision, recall, and F_1 -score curves of five classes respectively on the test set; (g) The precision-recall curves of five classes on the test set as well as their micro-averaging precision-recall curve; (h) The receiver operating characteristic (ROC) curves of five classes on the test set as well as their macro and micro-averaging ROC curves.

Table 4. The performance comparison of our proposed model with the previous methods in the partial agreement setting of the SCIAN dataset in terms of accuracy, precision, recall, specificity, and F₁-score metrics. Bold font shows the best results. All the metrics are described in percentages. The accuracy, precision, specificity, and F₁-score of the method in [21] were not reported directly, but calculated from its confusion matrix. The symbol ‘-’ stands for unreported results.

| Model | True Positive Rate | | | | | Accuracy (Weighted Average TPR) | Recall (Average TPR) | Precision (Macro) | Specificity (Macro) | F ₁ -Score (Macro) |
|---|--------------------|-----------|----------|-----------|-----------|---------------------------------------|----------------------------|----------------------|------------------------|----------------------------------|
| | Normal | Tapered | Pyriform | Amorphous | Small | | | | | |
| MorphoSpermGS (SVM with Zernike moments) [14] | 44 | 62 | 33 | 23 | 70 | 36 | 46 | - | - | - |
| MorphoSpermGS (SVM with Fourier descriptors) [14] | 57 | 68 | 53 | 15 | 54 | 34 | 49 | - | - | - |
| CE-SVM [19] | 62 | 64 | 50 | 30 | 82 | 44 | 58 | - | - | - |
| APDL [20] | 71 | 67 | 71 | 35 | 68 | 49 | 62 | - | - | - |
| FT-VGG [21] | 67 | 57 | 69 | 38 | 78 | 49 | 62 | 47 | 87 | 53 |
| Proposed model (MC-HSH) | 70 | 79 | 62 | 57 | 71 | 63 | 68 | 56 | 90 | 61 |

Table 5. The average confusion matrix on the stratified five-fold partition of the SCIAN dataset in the partial agreement setting, where each cell value (in percent) is the average of 15 runs (5 folds × 3 runs).

| | | | | | | |
|------------|-----------|-----------------|-----------|-----------|-----------|-----------|
| True Class | Normal | 70 | 3 | 3 | 20 | 4 |
| | Tapered | 2 | 79 | 5 | 13 | 1 |
| | Pyriform | 3 | 8 | 62 | 26 | 1 |
| | Amorphous | 10 | 16 | 8 | 57 | 9 |
| | Small | 10 | 2 | 1 | 16 | 71 |
| | | Normal | Tapered | Pyriform | Amorphous | Small |
| | | Predicted Class | | | | |

4.2. On the Stratified Five-Fold Partition of the SCIAN Dataset with the Total Agreement Setting

Our proposed model is further evaluated on the stratified five-fold partition of the SCIAN dataset in the total agreement setting. Similarly, we train the deep CNN architecture on each choice of training set in the total agreement setting. However, we no longer tune the hyperparameters since they have been tuned in the previous case of the partial agreement setting. The experimental results of our proposed model in the total agreement setting are shown in Figure 5a–h. The detailed experimental evaluations in the total agreement setting are available in Supplementary Materials (Figures S9–S16). Specifically, Figure 5a–b show typical classification accuracy and cost curves during the training on a specific choice of training and test sets. It is seen that our proposed model obtains very high classification accuracy in the total agreement setting when the training process converged. Our proposed model also attains a much higher accuracy and recall than the previous methods in the total agreement setting, which is clearly shown in Table 6 by simply comparing the precision, specificity, and F₁-score indices of our proposed model and the VGG model in [21]. For the elaborate comparisons with the models in [20,21], we employ the stratified five-fold cross-validation scheme in the total agreement setting. Figure 5c illustrates the confusion matrix of the classification for a typical run. We also compute the average confusion matrix over 15 runs (5 folds × 3 runs), as shown in Table 7. According to these tables, the Amorphous class remains the most difficult class to be differentiated from the other classes. Nevertheless, we can also find that the average TPRs of the Normal, Tapered, Pyriform, and Small classes become better. Therefore, the experimental results confirm that the Amorphous class is the most difficult to distinguish from the other classes. The precision, recall, and F₁-score curves of five classes on the test set in the total agreement setting through a typical run are shown in Figure 5d–f, respectively. From these three subfigures, we can again see that the five class curves of each of the precision, recall, and F₁-score globally tend to stabilize and increase as the number of epochs increases. We further plot the precision-recall curves of five classes on the test set as well as their micro-averaging precision-recall curve in Figure 5g for a typical run. It is clearly observed from this subfigure that the

Amorphous and Normal classes have a higher PR-AUC than the other classes, while the Pyriform class has the lowest one. Moreover, we plot the ROC curves of five classes on the test set as well as their macro and micro-averaging ROC curves in Figure 5h for a typical run. From this subfigure, we can see that the Normal class has the highest ROC-AUC, while the Amorphous class has the lowest one. The detailed results of each fold in the total agreement setting for each run are available in Table 9. From the last row of this table, we can also see low standard deviations of different indices from our proposed model with a training run in the total agreement setting, demonstrating the stability of our proposed model with the learning algorithm. As the agreement is strict in this case, common and essential features can be extracted effectively from the labeled images so that the classification results are improved considerably. In summary, our proposed model attains an overall accuracy of 77%, a macro precision of 64%, a macro recall of 88%, and a macro specificity of 94% in the total agreement setting, which are much better than the previous results.

Table 6. The performance comparison of our proposed model with the previous methods in the total agreement setting of the SCIAN dataset in terms of accuracy, precision, recall, specificity, and F₁-score metrics. The bold font shows the best results. All the metrics are described in percentages. The accuracy, precision, specificity, and F₁-score of the method in [21] were not reported directly, but calculated from its confusion matrix. The symbol ‘-’ stands for unreported results.

| Model | True Positive Rate | | | | | Accuracy (Weighted Average TPR) | Recall (Average TPR) | Precision (Macro) | Specificity (Macro) | F ₁ -Score (Macro) |
|----------------------------|--------------------|---------|----------|-----------|-------|---------------------------------------|----------------------------|----------------------|------------------------|----------------------------------|
| | Normal | Tapered | Pyriform | Amorphous | Small | | | | | |
| CE-SVM [19] | 74 | 70 | 92 | 30 | 100 | 46 | 73 | - | - | - |
| FT-VGG [21] | 72 | 67 | 95 | 44 | 84 | 53 | 72 | 45 | 90 | 55 |
| Proposed model (MC-HSH) | 80 | 86 | 100 | 72 | 100 | 77 | 88 | 64 | 94 | 74 |

Table 7. The average confusion matrix on the stratified five-fold partition of the SCIAN dataset in the total agreement setting, where each cell value (in percent) is the average of 15 runs (5 folds × 3 runs).

| | | | | | | |
|-----------------|-----------|--------|---------|----------|-----------|-------|
| True Class | Normal | 80 | 0 | 3 | 10 | 7 |
| | Tapered | 2 | 86 | 0 | 9 | 3 |
| | Pyriform | 0 | 0 | 100 | 0 | 0 |
| | Amorphous | 8 | 11 | 2 | 72 | 7 |
| | Small | 0 | 0 | 0 | 0 | 100 |
| | | Normal | Tapered | Pyriform | Amorphous | Small |
| Predicted Class | | | | | | |

Table 8. The stratified five-fold cross-validation results for the morphological classification of human sperm heads in the partial agreement setting of the SCIAN dataset for every fold and every run, where all the metrics except Matthews correlation coeffi

Table 9. The stratified five-fold cross-validation results for the morphological classification of human sperm heads in the total agreement setting of the SCIAN dataset for every fold and every run, where all the metrics except Matthews correlation coefficient (MCC) and Cohen’s kappa score (CKS) are described in percentages.

| Fold | Run | Precision | | Recall | | Specificity | | F ₁ -Score | | Jaccard | | G-mean | | ROC-AUC | | | MCC | CKS |
|--------------------|--------|-----------|----------|--------|-------------------|-------------|----------|-----------------------|----------|---------|----------|--------|----------|---------|--------|--------|--------|--------|
| | | Macro | Weighted | Macro | Weighted Accuracy | Macro | Weighted | Macro | Weighted | Macro | Weighted | Macro | Weighted | Macro | Micro | Micro | | |
| 1 | First | 69 | 88 | 82 | 73 | 94 | 97 | 68 | 77 | 57 | 64 | 87 | 84 | 95 | 94 | 77 | +0.61 | +0.56 |
| | Second | 62 | 85 | 88 | 70 | 93 | 96 | 67 | 72 | 52 | 57 | 90 | 81 | 96 | 93 | 75 | +0.60 | +0.54 |
| | Third | 62 | 87 | 83 | 75 | 94 | 96 | 66 | 79 | 52 | 67 | 88 | 85 | 94 | 93 | 75 | +0.62 | +0.59 |
| 2 | First | 69 | 82 | 83 | 78 | 93 | 85 | 72 | 79 | 61 | 67 | 87 | 81 | 94 | 95 | 82 | +0.61 | +0.60 |
| | Second | 60 | 84 | 87 | 77 | 93 | 90 | 65 | 79 | 50 | 66 | 89 | 83 | 94 | 94 | 81 | +0.62 | +0.60 |
| | Third | 62 | 80 | 83 | 77 | 92 | 94 | 68 | 77 | 52 | 64 | 87 | 80 | 95 | 95 | 85 | +0.59 | +0.58 |
| 3 | First | 62 | 83 | 91 | 76 | 92 | 91 | 71 | 77 | 56 | 63 | 92 | 83 | 95 | 94 | 77 | +0.63 | +0.60 |
| | Second | 64 | 84 | 87 | 80 | 93 | 87 | 72 | 81 | 57 | 69 | 90 | 83 | 95 | 94 | 79 | +0.65 | +0.64 |
| | Third | 57 | 84 | 89 | 74 | 94 | 94 | 64 | 75 | 49 | 61 | 91 | 83 | 94 | 93 | 73 | +0.61 | +0.57 |
| 4 | First | 60 | 87 | 89 | 79 | 95 | 95 | 67 | 81 | 52 | 69 | 92 | 87 | 97 | 96 | 86 | +0.66 | +0.64 |
| | Second | 70 | 81 | 85 | 77 | 92 | 86 | 75 | 78 | 63 | 65 | 89 | 81 | 95 | 95 | 82 | +0.59 | +0.58 |
| | Third | 66 | 86 | 91 | 79 | 95 | 92 | 74 | 80 | 59 | 68 | 92 | 85 | 97 | 95 | 82 | +0.66 | +0.64 |
| 5 | First | 59 | 86 | 92 | 77 | 94 | 95 | 68 | 78 | 53 | 64 | 93 | 85 | 97 | 96 | 84 | +0.66 | +0.62 |
| | Second | 67 | 89 | 94 | 79 | 95 | 97 | 75 | 80 | 61 | 68 | 94 | 87 | 98 | 96 | 86 | +0.70 | +0.66 |
| | Third | 67 | 87 | 94 | 82 | 96 | 95 | 76 | 83 | 61 | 71 | 94 | 88 | 98 | 96 | 86 | +0.71 | +0.69 |
| Average | | 64 | 85 | 88 | 77 | 94 | 93 | 70 | 78 | 56 | 66 | 90 | 84 | 96 | 95 | 81 | + 0.63 | +0.61 |
| Standard deviation | | 0.0404 | 0.0261 | 0.0405 | 0.0300 | 0.0123 | 0.0401 | 0.0394 | 0.0267 | 0.0456 | 0.0356 | 0.0247 | 0.0243 | 0.0145 | 0.0112 | 0.0442 | 0.0374 | 0.0421 |

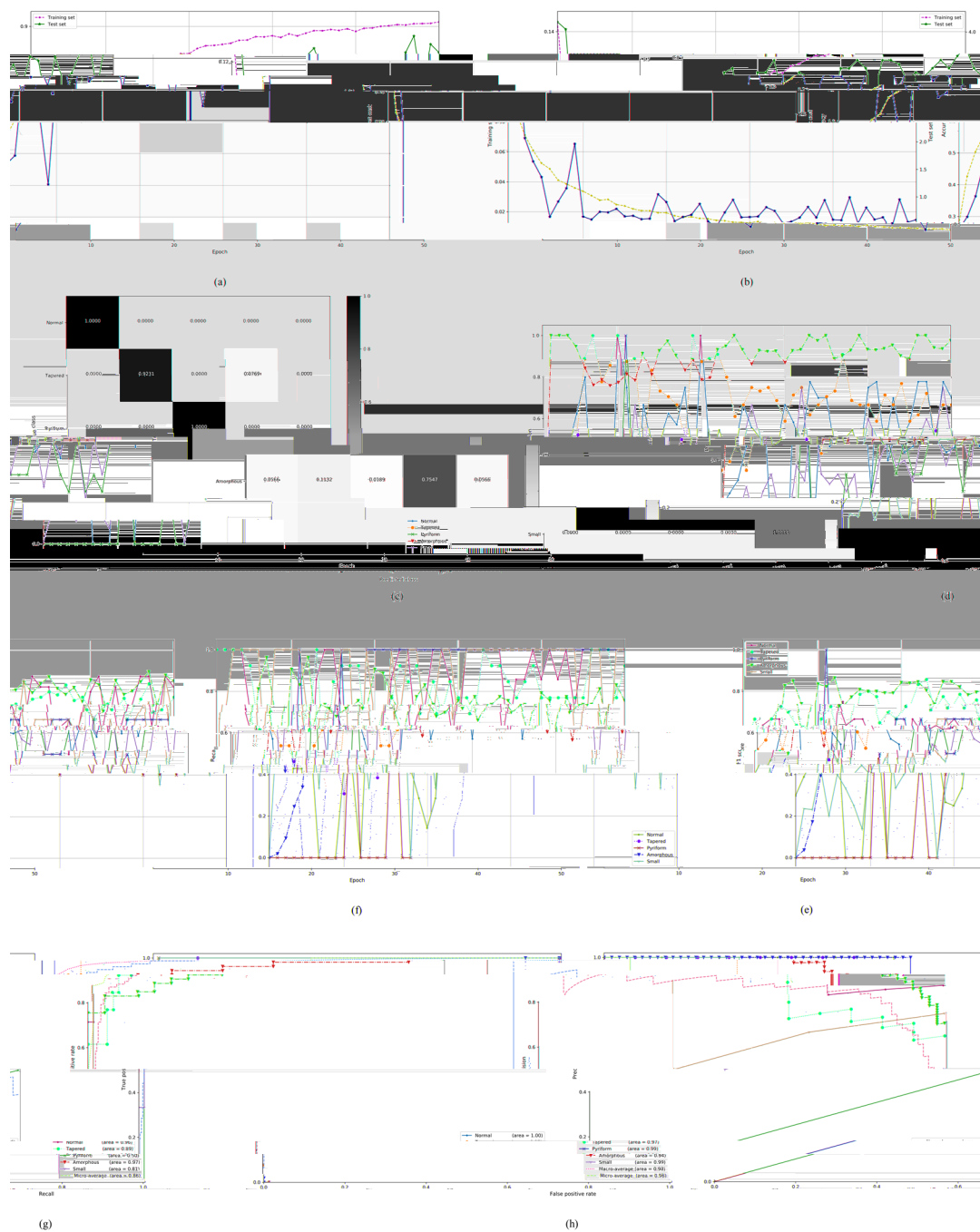


Figure 5. The experimental results in the total agreement setting of the SCIAN dataset: (a,b) Typical classification accuracy and cost curves with the number of epochs during the training on the training and test sets; (c) The confusion matrix on the test set; (d–f) The precision, recall, and F₁-score curves of five classes respectively the on test set; (g) The precision-recall curves of five classes on the test set as well as their micro-averaging precision-recall curve; (h) The receiver operating characteristic (ROC) curves of five classes on the test set as well as their macro and micro-averaging ROC curves.

4.3. On the Stratified Five-Fold Partition of the HuSHeM Dataset

Our proposed model is finally evaluated on the HuSHeM [20] dataset. This is another dataset for the morphological classification of human sperm heads with 216 images (54 Normal, 53 Tapered, 57 Pyriform, and 52 Amorphous). Its images are also manually annotated by three human experts, but only the images with three-expert agreement are recorded. Each image consists of 131 by 131 pixels, being taken at 100× magnification.

In the pre-processing step, we first rotate the images so that all human sperm heads share the same orientation. We then crop the sample images so that the sperm heads appear in the center of the images. After this step, the images are reduced to 90 by 90 pixels. Approximately 80% of the images are considered for training and the remaining images for the evaluation. We further employ data augmentation techniques to solve the scarcity of training images. As for the data augmentation, we adopt three common techniques as we used in the training set of the SCIAN dataset. For rotation, we rotate the training image by -25 to 25 degrees. For translation, we shift $\sim 6\%$ of the original image to the left, the right, up, and down. For flipping, we vertically flip the image. Due to the same distribution of classes within this dataset, we apply equal augmentation options to each class. For the convenience of comparison, we also adopt a stratified five-fold cross-validation scheme as used in [20,21]. We utilize the Adam learning algorithm to train our proposed deep CNN model with mini a batch size of 256 for 25 epochs for the HuSHeM dataset. To compare the performance of our proposed model directly with previously published results [20,21], each five-fold cross-validation procedure runs three times for stability.

The experimental results of our proposed model on the HuSHeM dataset are shown in Figure 6a–e. The detailed experimental results are shown in Supplementary Materials (Figures S17–S21). The experimental results of our proposed model as well as the previous methods on the HuSHeM dataset are shown in Table 10. It is clearly seen that our proposed model achieves better accuracy, recall, precision, specificity, and F_1 -score than previous methods. Moreover, from the confusion matrix of our proposed model (Table 11), we can see that Pyriform classes in the test set are predicted 97% correctly. Results on the HuSHeM dataset are the average of 15 runs (5 folds \times 3 runs). We also plot the precision-recall curves of four classes on the test set as well as their micro-averaging precision-recall curve in Figure 6d for a typical run. Furthermore, we plot the ROC curves of four classes on the test set as well as their macro and micro-averaging ROC curves in Figure 6e for a typical run. Finally, we summarize the detailed results of each fold for each run in Table 12 which includes all the possible evaluation metrics such as the precision, recall, specificity, F_1 -score, Jaccard, G-mean, ROC-AUC, PR-AUC, MCC, CKS, and evaluation time.

Table 10. The performance comparison of our proposed model with the previous methods on the HuSHeM dataset in terms of accuracy, recall, precision, specificity, and F_1 -score metrics. Bold font shows the best results. All the metrics are described in percentages. The specificity of the methods in [19–21] were not reported directly, but calculated from their confusion matrices.

| Model | True Positive Rate | | | | Accuracy | Recall | Precision | Specificity | F_1 -Score |
|-------------------------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|
| | Normal | Tapered | Pyriform | Amorphous | | | | | |
| CE-SVM [19] | 75.9 | 77.3 | 85.9 | 75.0 | 78.5 | 78.5 | 80.5 | 92.9 | 78.9 |
| APDL [20] | 94.4 | 94.3 | 87.7 | 94.2 | 92.2 | 92.3 | 93.5 | 97.5 | 92.9 |
| FT-VGG [21] | 96.4 | 94.5 | 92.3 | 93.2 | 94.0 | 94.1 | 94.7 | 98.1 | 94.1 |
| Proposed model (MC-HSH) | 95.8 | 94.5 | 96.6 | 96.4 | 95.7 | 95.5 | 96.1 | 98.5 | 95.5 |

Table 11. The average confusion matrix of our proposed model on the HuSHeM dataset, where each cell value (in percent) is the average of 15 runs.

| | | | | | |
|------------|-----------|-----------------|-----------|-----------|-----------|
| True Class | Normal | 96 | 3 | 1 | 0 |
| | Tapered | 1 | 94 | 2 | 3 |
| | Pyriform | 0 | 1 | 97 | 2 |
| | Amorphous | 2 | 2 | 0 | 96 |
| | | Normal | Tapered | Pyriform | Amorphous |
| | | Predicted Class | | | |

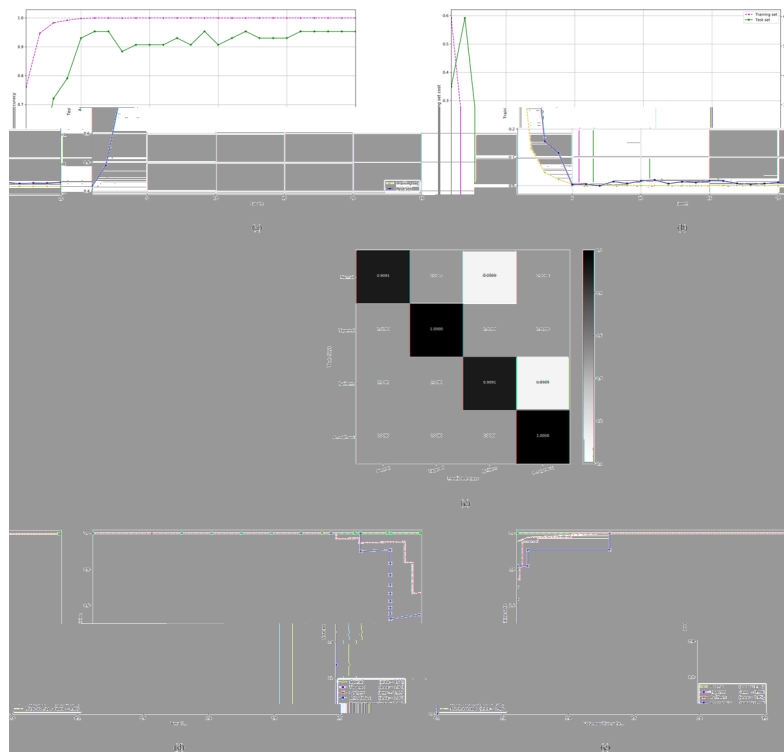


Figure 6. The experimental results of the HuSHeM dataset: (a,b) Typical classification accuracy and cost curves with the number of epochs during the training on the training and test sets; (c) The confusion matrix on the test set; (d) The precision-recall curves of four classes on the test set as well as their micro-averaging precision-recall curve; (e) The receiver operating characteristic (ROC) curves of four classes on the test set as well as their macro and micro-averaging ROC curves.

Table 12. The stratified five-fold cross-validation results for the morphological classification of human sperm heads on the HuSHeM dataset for every fold and every run, where all the metrics except Matthews correlation coefficient (MCC) and Cohen’s kappa score (CKS) are described in percentages.

| Fold | Run | Precision | | Recall | | Specificity | | F ₁ -Score | | Jaccard | | G-mean | | ROC-AUC | | PR-AUC | MCC | CKS | Evaluation Time per Image (milliseconds) |
|--------------------|--------|-----------|----------|--------|-------------------|-------------|----------|-----------------------|----------|---------|----------|--------|----------|---------|--------|--------|--------|--------|--|
| | | Macro | Weighted | Macro | Weighted Accuracy | Macro | Weighted | Macro | Weighted | Macro | Weighted | Macro | Weighted | Macro | Micro | Micro | | | |
| 1 | First | 98 | 98 | 98 | 98 | 99 | 99 | 98 | 98 | 96 | 96 | 98 | 98 | 100 | 100 | 100 | +0.97 | +0.97 | ~0.9 |
| | Second | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | +1.00 | +1.00 | |
| | Third | 98 | 98 | 98 | 98 | 99 | 99 | 98 | 98 | 96 | 96 | 98 | 98 | 100 | 100 | 100 | +0.97 | +0.97 | |
| 2 | First | 95 | 96 | 95 | 95 | 98 | 98 | 95 | 95 | 91 | 91 | 97 | 97 | 100 | 99 | 98 | +0.94 | +0.94 | |
| | Second | 95 | 96 | 95 | 95 | 98 | 98 | 95 | 95 | 91 | 91 | 97 | 97 | 100 | 100 | 99 | +0.94 | +0.94 | |
| | Third | 93 | 93 | 93 | 93 | 98 | 98 | 93 | 93 | 87 | 87 | 95 | 95 | 100 | 99 | 99 | +0.91 | +0.91 | |
| 3 | First | 95 | 96 | 95 | 95 | 98 | 98 | 95 | 95 | 91 | 91 | 97 | 97 | 100 | 99 | 99 | +0.94 | +0.94 | |
| | Second | 95 | 96 | 95 | 95 | 98 | 98 | 95 | 95 | 91 | 91 | 97 | 97 | 99 | 99 | 98 | +0.94 | +0.94 | |
| | Third | 95 | 96 | 95 | 95 | 98 | 98 | 95 | 95 | 91 | 91 | 97 | 97 | 100 | 99 | 98 | +0.94 | +0.94 | |
| 4 | First | 98 | 98 | 98 | 98 | 99 | 99 | 98 | 98 | 95 | 96 | 98 | 98 | 100 | 100 | 100 | +0.97 | +0.97 | |
| | Second | 95 | 95 | 95 | 95 | 99 | 99 | 95 | 95 | 91 | 91 | 97 | 97 | 100 | 100 | 100 | +0.94 | +0.94 | |
| | Third | 95 | 96 | 95 | 95 | 99 | 99 | 95 | 95 | 91 | 92 | 96 | 97 | 100 | 100 | 100 | +0.94 | +0.94 | |
| 5 | First | 94 | 94 | 93 | 93 | 98 | 98 | 93 | 93 | 87 | 87 | 95 | 95 | 100 | 100 | 100 | +0.91 | +0.91 | |
| | Second | 95 | 96 | 96 | 95 | 99 | 99 | 95 | 95 | 91 | 91 | 97 | 97 | 100 | 99 | 98 | +0.94 | +0.94 | |
| | Third | 94 | 94 | 93 | 93 | 98 | 98 | 93 | 93 | 87 | 87 | 95 | 95 | 100 | 99 | 99 | +0.91 | +0.91 | |
| Average | | 96 | 96 | 96 | 96 | 99 | 99 | 96 | 96 | 92 | 92 | 97 | 97 | 100 | 100 | 99 | +0.94 | +0.94 | |
| Standard deviation | | 0.0191 | 0.0181 | 0.0207 | 0.0207 | 0.0064 | 0.0064 | 0.0207 | 0.0207 | 0.0365 | 0.0372 | 0.0133 | 0.0131 | 0.0026 | 0.0052 | 0.0086 | 0.0250 | 0.0250 | |

5. Discussion and Conclusions

We have established an advanced deep CNN architecture, MC-HSH, specially for the morphological classification of human sperm heads. In this deep learning architecture, there are a total of 53 convolutional layers. Before each convolutional layer, the batch normalization and LeakyReLU are used. We also apply the channel-wise concatenation and element-wise addition to make this model more effective for the morphological classification of human sperm heads. We employ the L_2 penalty as the kernel regularizer in the dense layer to prevent overfitting. We utilize several layers and multiple filter sizes, but fewer filters and parameters, and we also make a new arrangement of convolutional layers, addition and concatenation operations for this classification task.

According to the WHO criteria [9], human sperm heads are classified into categories such as Normal, Tapered, Pyriform, Amorphous, and Small and their morphological classification is very challenging. Based on a golden standard SCIAN dataset of microscopic sperm images and the HuSHeM dataset, data-driven machine learning models and algorithms can be utilized to solve this difficult problem. By making careful pre-processing, partition, and argumentation of the SCIAN and the HuSHeM datasets, we design a specialized deep CNN architecture for the morphological classification of sperm heads based on the microscopic human sperm head images. The stratified five-fold cross-validation results demonstrate that our proposed model (along with the deep learning algorithm) is much more effective than the previous methods [14,19–21] for the morphological classification of human sperm heads. The performance indices on five classes (see Tables 4–7) indicate that it is reliable in recognizing the images in the Normal class as well as the four abnormal classes. By attaining the embryologist level performance of the classification, our proposed model is also a balanced classifier where the TPR is similar to the positive predictive value (PPV).

It can be found from Tables 4 and 6 that the previous methods are not so powerful to extract effective features from microscopic images for the classification of human sperm heads. Our proposed model achieves 68% and 88% average TPR on the SCIAN dataset in the partial and total agreement settings, respectively. We find out that our proposed model improves the accuracy and recall by a factor of 29% and 10%, respectively, in the partial agreement setting and 46% and 22%, respectively, in the total agreement setting compared with the state-of-the-art results reported in [21]. In the total agreement setting, our proposed model achieves a much better accuracy (77%) and recall (88%) because the training set has more images and the test set has the total expert agreement images in comparison with the accuracy (63%) and recall (68%) of the partial agreement setting. Our proposed model can extract the morphometric features for seminogram which are significant for sperm binding to the oocyte. The morphological classification of human sperm heads is an intricate problem because of intrinsic inter-class similarities and intra-class variabilities. Our proposed model achieves better classification results than the previous state-of-the-art methods without using transfer learning. On the HuSHeM dataset, the results of our proposed model are also better than the state-of-the-art results. Our proposed approach achieves 96% accuracy and 95% recall on the HuSHeM dataset. The accuracy, recall, precision, and F_1 -score increase approximately 2%, whereas the specificity improves roughly 0.5% in comparison with [21]. The results of our proposed model are much better on the HuSHeM dataset than the SCIAN dataset. This improvement is due to three main reasons: (1) the HuSHeM dataset has only four sperm head classes; (2) its images have a high resolution; (3) and all of its images are 3-out-of-3 human expert agreement. The evaluation time of our proposed model is ~0.2 milliseconds (ms) for the SCIAN dataset, while ~0.9 ms for the HuSHeM dataset per image.

Developing an automated classification system of human sperm heads can greatly reduce the workload of embryologists and also decrease the subjectivity and inaccuracy of the classification induced by the human error. This automated system can become necessary and more valuable when experienced embryologists are not readily available and for inexperienced clinicians in underdeveloped countries. In fact, the classification results of our proposed model are comparable to those of the domain experts. Consequently, our proposed model can even be used to assign a class label to any new sperm head image, and this deep CNN architecture is good to expedite the automatic classification

procedure of human sperm heads. Indeed, our research provides more strong evidence that the deep learning approach is able to play a key role in healthcare systems, assisting doctors to achieve higher conception and gestation rates. Our proposed architecture shows the potential of deep learning to surpass embryologists in terms of throughput, accuracy and reliability.

It is worth indicating the limitations of this study. As mentioned before, experiments are conducted on two publicly available datasets. The SCIAN dataset has 1132 and 384 human sperm heads images in the partial and total agreement settings, respectively, while the HuSHeM dataset has only 216 human sperm head images. These numbers of images are relatively small. Consequently, to obtain better generalizability, it is essential to increase the number of images for experimentation in the future. Secondly, due to limited computational power and memory the training time is high. Lastly, additional work remains to be done to evaluate the deep learning models in fertility clinics.

Supplementary Materials: Available online at <http://www.mdpi.com/2075-4418/10/5/325/s1>. Figure S1. Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 classification accuracy curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds \times 3 runs) are illustrated; Figure S2. Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 cost curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds \times 3 runs) are illustrated; Figure S3. Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 confusion matrixes on each of five possible choices of the test sets for 3 runs (5 folds \times 3 runs) are illustrated; Figure S4. Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 precision curves of each class with the number of epochs during the training on

epochs of the HuSHeM dataset, where 15 cost curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds \times 3 runs) are illustrated; Figure S19. Detailed experimental results of the proposed model through 25 epochs of the HuSHeM dataset, where 15 confusion matrixes on each of five possible choices of the test sets for 3 runs (5 folds \times 3 runs) are illustrated; Figure S20. Detailed experimental results of proposed model of the HuSHeM dataset, where 15 precision-recall curves of each class and their micro-averaging precision-recall curve on each of five possible choices of the test sets for 3 runs (5 folds \times 3 runs) are illustrated. A high area under the curve signifies the high precision as well as high recall; Figure S21. Detailed experimental results of proposed model of the HuSHeM dataset, where 15 receiver operating characteristic (ROC) curves of each class and their macro and micro-averaging ROC curves on each of five possible choices of the test sets for 3 runs (5 folds \times 3 runs) are illustrated. These plots show the tradeoff between the true positive rate and false positive rate.

Author Contributions: I.I. and J.M. devised the project, the main conceptual ideas and proof outline; I.I. designed and performed the experiments; I.I. drafted the manuscript, designed the figures and tables with support from G.M.; I.I. and J.M. contributed to the interpretation of the results; J.M. supervised the project; I.I., J.M., and G.M. reviewed and revised the manuscript prior to submission. All authors have read and approved the final submitted manuscript.

Funding: This work was supported by the National Key Research and Development Program of China under grant 2018AAA0100205.

Acknowledgments: The authors would like to thank Zhe Sage Chen, Zhongyu Joan Lu and Qiang Xu for their kind help.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Villarreal, M.R. Complete Diagram of a Human Spermatozoon. Available online: https://commons.wikimedia.org/wiki/File:Complete_diagram_of_a_human_spermatozoa_en.svg (accessed on 9 October 2019).
2. Zegers-Hochschild, F.; Adamson, G.D.; De Mouzon, J.; Ishihara, O.; Mansour, R.; Nygren, K.; Sullivan, E.; Vanderpoel, S. International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) revised glossary of ART. *Fertil. Steril.* **2009**, *92*, 1520–1524. [[CrossRef](#)] [[PubMed](#)]
3. Blasco, V.; Pinto, F.M.; Gonz, C.; Santamar, E.; Candenias, L.; Fern, M. Tachykinins and Kisspeptins in the Regulation of Human Male Fertility. *J. Clin. Med.* **2020**, *9*, 113. [[CrossRef](#)] [[PubMed](#)]
4. Wang, S.-C.; Wang, S.-C.; Li, C.-J.; Lin, C.-H.; Huang, H.-L.; Tsai, L.-M.; Chang, C.-H. The Therapeutic Effects of Traditional Chinese Medicine for Poor Semen Quality in Infertile Males. *J. Clin. Med.* **2018**, *7*, 239. [[CrossRef](#)] [[PubMed](#)]
5. Kidd, S.A.; Eskenazi, B.; Wyrobek, A.J. Effects of male age on semen quality and fertility: A review of the literature. *Fertil. Steril.* **2001**, *75*, 237–248. [[CrossRef](#)]
6. Barone, M.A.; Roelke, M.E.; Howard, J.; Brown, J.L.; Anderson, A.E.; Wildt, D.E. Reproductive Characteristics of Male Florida Panthers: Comparative studies from Florida, Texas, Colorado, Latin America, and North American Zoos. *J. Mammal.* **1994**, *75*, 150–162. [[CrossRef](#)]
7. Monte, G.L.; Murisier, F.; Piva, I.; Germond, M.; Marci, R. Focus on intracytoplasmic morphologically selected sperm injection (IMSI): A mini-review.

15. Yi, W.J.; Park, K.S.; Paick, J.S. Parameterized characterization of elliptic sperm heads using Fourier representation and wavelet transform. In Proceedings of the Annual International Conference of the ZEEE Engineering in Medicine and Biology Society, Hong Kong, China, 1 November 1998; Volume 20, pp. 974–977.
16. Li, J.; Tseng, K.K.; Dong, H.; Li, Y.; Zhao, M.; Ding, M. Human sperm health diagnosis with principal component analysis and k-nearest neighbor algorithm. In Proceedings of the International Conference on Medical Biometrics, Shenzhen, China, 30 May–1 June 2014; IEEE: Shenzhen, China, 2014; pp. 108–113.
17. Beletti, M.E.; Costa, L.D.F.; Viana, M.P. A comparison of morphometric characteristics of sperm from fertile *Bos taurus* and *Bos indicus* bulls in Brazil. *Anim. Reprod. Sci.* **2005**, *85*, 105–116. [[CrossRef](#)] [[PubMed](#)]
18. Severa, L.; Máchal, L.; Švábová, L.; Mamica, O. Evaluation of shape variability of stallion sperm heads by means of image analysis and Fourier descriptors. *Anim. Reprod. Sci.* **2010**, *119*, 50–55. [[CrossRef](#)]
19. Chang, V.; Heutte, L.; Petitjean, C.; Hartel, S.; Hitschfeld, N. Automatic classification of human sperm head morphology. *Comput. Biol. Med.* **2017**, *84*, 205–216. [[CrossRef](#)]
20. Shaker, F.; Monadjemi, S.A.; Alirezaie, J.; Naghsh-Nilchi, A.R. A dictionary learning approach for human sperm heads classification. *Comput. Biol. Med.* **2017**, *91*, 181–190. [[CrossRef](#)]
21. Riordon, J.; McCallum, C.; Sinton, D. Deep learning for the classification of human sperm. *Comput. Biol. Med.* **2019**, *111*. [[CrossRef](#)]
22. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Jan, C.V.; Krause, J.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
23. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
24. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Volume 9351, pp. 234–241.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
26. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
27. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
28. Iqbal, I.; Shahzad, G.; Rafiq, N.; Mustafa, G.; Ma, J. Deep learning-based automated detection of human knee joint's synovial fluid from magnetic resonance images with transfer learning. *IET Image Process.* **2020**. (In press)
29. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
30. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16 June–21 June 2013; Volume 28.
31. LeCun, Y.A.; Bottou, L.; Orr, G.B.; Müller, K.-R. *Efficient BackProp*; Springer: Heidelberg, Germany, 2012.
32. Kingma, D.P.; Ba, J.L. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
33. Chollet, F. Keras (2015). Available online: <http://keras.io> (accessed on 9 October 2019).
34. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, arXiv:1603.04467.

